International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Emotion Detection using Twitter Datasets and Spacy Algorithm

Dr. P Meena Kumari[1],Kotla Srinija Reddy[2], Julakanti Nikhitha[3], Alakanti Sai Charan[4], B Akshay Kumar[5]
[2,3,4,5] UG Scholars, Department of CSE, **AVN Institute of Engineering and Technology,**Hyderabad, Telangana, India.
[1]Assoiate Professor, Department of CSE, **AVN Institute of Engineering and Technology**, Hyderabad, Telangana, India.

ABSTRACT :

People show emotions for everyday communication. Emotions are identified by facial expressions, behavior, writing, speaking, gestures and physical actions. Emotion plays a vital role in the interaction between two people. The detection of emotions through text is a challenge for researchers. Emotion detection from the text can be useful for real-world application. Automatic emotion detection in the original text aims to recognize emotions in any digital medium by using natural language processing techniques and different approaches. Enabling machines with the ability to recognize emotions in a particular kind of text such as twitter's tweet has important applications in sentiment analysis and affective computing. We have worked on the newly published gold dataset (AIT-2018) and propose a model consisting of lexicalbased using WordNet-Affect and EmoSenticNet with supervised classifiers for detecting emotions in a tweet text**.**

INTRODUCTION :

Language is known to be a powerful instrument for communicating and conveying information and for expressing emotions. Currently, neuroscience, psychology, cognitive sciences, computer sciences, and computational sciences are studying emotional identification widely. The integration into our everyday life of several interactive online diaries, journals, and individual blogs helps meet important social-interaction needs [1].

In the world of today's social networks, users share their opinions and emotions in their way through different medium like Twitter, Instagram, Facebook, and many more. Where millions, in their everyday lives express their views and opinions and also their emotions on or about a particular thing through social networks [2]. This gave the researchers an excellent opportunity to analyze the emotions of social networking users' activities. These large numbers of data, generated by social networks contain feelings, opinions, and emotions of people from day to day. Different emotional analytical research on the social platform has been underway over the years. As the public have different thoughts, it becomes a challenge to analyze the correct emotion from social data. This makes it clear about the need to work on these problems and it offers many possibilities for future research into the hidden identification of emotions of users in general or emotions of users on a specific topic, etc.

Here we will study and analyze previous works done in this area, identify research scope, understand the process, methods used and finally propose a model that will help us to detect an emotion which is expressed in tweets. We will work on AIT-2018 dataset [3] and our proposed methodology consists of different phases, a basic idea of the whole model is shown in figure 1 and detailed worked is described in the following sections.

LITERATURE SURVEY :

1.A Survey On Emotion Detection Techniques using Text in Blogposts :

Emotion can be expressed in many ways that can be seen such as facial expression and gestures, speech and by written text. Emotion Detection in text documents is essentially a content - based classification problem involving concepts from the domains of Natural Language Processing as well as Machine Learning. In this paper emotion recognition based on textual data and the techniques used in emotion detection are discussed.

2. The Impact of Social Media on Intercultural Adaptation :

Social media has become increasingly popular components of our everyday life in today's globalizing society. It provides a context where people across the world can communicate, exchange messages, share knowledge, and interact with each other regardless of the distance that separates them. Intercultural adaptation involves the process of promoting understanding through interaction to increase the level of fitness so that the demands of a new cultural environment can be met. Research shows that people tend to use social media to become more integrated into the host culture during their adaptation and to maintain connections to their home countries. This paper attempts to investigate the impact of using social media on the intercultural adaptation process. In-depth interviews of international students of a U.S. university are conducted. Based on the results of the analysis, directions for future studies in this line of research are also discussed.

3. Use of Word Clustering to Improve Emotion Recognition from Short Tex:

Emotion recognition is an important component of affective computing, and is significant in the implementation of natural and friendly human-computer interaction. An effective approach to recognizing emotion from text is based on a machine learning technique, which deals with emotion recognition as a classification problem. However, in emotion recognition, the texts involved are usually very short, leaving a very large, sparse feature space, which decreases the performance of emotion classification. This paper proposes to resolve the problem of feature sparseness, and largely improve the emotion recognition performance from short texts by doing the following: representing short texts with word cluster features, offering a novel word clustering algorithm, and using a new feature weighting scheme. Emotion classification experiments were performed with different features and weighting schemes on a publicly available dataset. The experimental results suggest that the word cluster features and the proposed weighting scheme can partly resolve problems with feature sparseness and emotion recognition performance. © 2016. The Korean Institute of Information Scientists and Engineers.

4. Multiclass Emotion Extraction from Sentences :

This paper aims to investigate the extraction of different classes of emotion from sentences using supervised machine learning technique, Multinomial Naïve Bayes (MNB). Here a bag of word approach is used to capture the emotions. The unigrams are mainly used for this and the bigrams and trigrams are used to capture lower order dependencies. The work is done on the ISEAR dataset [14]. The experiments with different feature sets selected using Weighted log-likelihood score (WLLS) [12] shows that the MNB classifier provides good results when the unigram feature set size is 450 which provides an average accuracy of 76.96% across all emotion classes.

5. A Hybrid Model for Automatic Emotion Recognition in Suicide Notes :

We describe the Open University team's submission to the 2011 i2b2/VA/Cincinnati Medical Natural Language Processing Challenge, Track 2 Shared Task for sentiment analysis in suicide notes. This Shared Task focused on the development of automatic systems that identify, at the sentence level, affective text of 15 specific emotions from suicide notes. We propose a hybrid model that incorporates a number of natural language processing techniques, including lexicon-based keyword spotting, CRF-based emotion cue identification, and machine learning-based emotion classification. The results generated by different techniques are integrated using different vote-based merging strategies. The automated system performed well against the manually-annotated gold standard, and achieved encouraging results with a micro-averaged F-measure score of 61.39% in textual emotion recognition, which was ranked 1st place out of 24 participant teams in this challenge. The results demonstrate that effective emotion recognition by an automated system is possible when a large annotated corpus is available.

6.Computational approaches for emotion detection in text :

Emotions are part and parcel of human life and among other things, highly influence decision making. Computers have been used for decision making for quite some time now but have traditionally relied on factual information. Recently, interest has been growing among researchers to find ways of detecting subjective information used in blogs and other online social media. This paper presents emotion theories that provide a basis for emotion models. It shows how these models have been used by discussing computational approaches to emotion detection. We propose a hybrid based architecture for emotion detection. The SVM algorithm is used for validating the proposed

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

architecture and achieves a prediction accuracy of 96.43% on web blog data.

EXITING SYSTEM :

There has been many works in this area for the last couple of years. In this section we are going to view some of the previous works done by different authors. In [4] the authors created a corpus of Twitter tweets and used corpus annotation study to prepare an annotated corpus. Multi-class SVM kernels were used for learning model. For features selection Unigrams, Bigrams, Personal, pronouns, adjectives, Word-net Affect emotion lexicon, Word-net Affect emotion and Dependency–parsing features. In [5] the authors first fetched the tweets from Twitter to create a dataset. Then they obtain target based extended features model. They trained four different supervised classifiers, Naïve Bayes (NB), Support Vector Machine (SVM), Maximum Entropy (MaxEn), Artificial Neural Networks (ANN). SVM combined with Principal Component Analysis (PCA) obtains the maximum accuracy. In [6] at first, the authors preprocessed the training dataset and took similarity measurements amongst the data. Then using semantic similarity all the emotion labeled corpus are clustered. In the training phase, the authors represented each text as a feature vector and the SVM learning algorithm is applied to train an emotion classifier. In [7] the authors focus on identifying seven different classes of emotions - Anger, Disgust, Fear, Guilt, Joy, Sadness, Shame. To extract features, preprocessed data is tokenized and then stemmed using porter stemming algorithm. Authors used Unigram, Bigram and Trigram feature. WLLS (Weighted Log-likelihood Score) scheme is applied to score n-grams with respect to each emotion resulting in a feature vector table. In the method, authors used MNB (Multinomial Naïve Bayes) as a classifier which is trained by the top scored n-grams and accuracy tested with different feature sets. In [8] the author showed a hybrid model for emotion detection. In this model, it contains lexiconkeyword spotting, CRF based emotion detection using NB, MaxEn, and SVM. In [9] the authors have used a Hidden Markov Model which determines the emotion of the text. They considered each sentence contains many sub ideas and each idea is considered as an event that might cause a transition of a state. In [10] the author created an automatic emotion detection system which can identify emotions in tweets streams. His approach included two-part, training an offline emotion classifier model which is based on his work from [11] and in the second part he performed a two-step classification to identify tweets containing emotions and classifying these tweets into a more fine-grained category using soft classification techniques. In [12] the authors tried to classify comments regarding a specific crisis on social media. They used the emotion of anger considering the fact that this same technique can be applied for other emotions as well. They performed a short survey collecting 1192 responses in which the people are requested to comment under a news headline using social media. Using this as training set they obtain an accuracy of 90% in classifying anger in their dataset. They used logistic regression coefficients to select their features and random forest as their main classifier.

EMOTION DETECTION FROM TWEET :

For emotion detection, there are four kinds of text-based techniques as follows keyword spotting method, lexical affinity method, learning based method and hybrid methods [13]. For detecting emotions from tweets, we have used the lexical affinity method combined with learning-based methods to automatically classify multi-class emotions from our dataset. We have used WordNet-Affect [14] and EmoSenticNet [15] emotion lexicon to extract the emotion containing words as features from the tweet separately. WordNet-Affect returns the emotion representing words from the tweets which is then considered as features but in most cases, it is unable to take the words which may not be an emotion word but do represent an emotion. For a small set of words WordNetAffect can determine if the word represents one of the six basic emotions. The main drawback of WordNet-Affect is that it can not give an intensity for the words as some words though they are a synonym of each other may represent different type of emotion with respect to text. On the other hand EmoSenticNet is an extension of WordNet-Affect which also apply the SenticNet [16] rules. It then also finds the features which are not contained in WordNet-Affect.

Then we have used term frequency and inverse document frequency on the features to give the emotion features a better score, after that, we have used some different supervised algorithm for emotion classification. We have used Naïve Bayesian, Decision Tree and Support Vector Machine for emotion classification, all of these are supervised machine learning algorithm.

We have tried to experiment ourselves with detecting the emotion from a text document. Below we are presenting our proposed methodology in figure 1. The following sections describe each process in details.

DATASET :

For our dataset we have taken SemEval-2018 Affect in Tweets Distant Supervision Corpus (AIT-2018 Dataset). Using twitter API these tweets are crawled from twitter from tweets that included emotion-related words such as ’#angry’, ’#annoyed’, ’#panic’, ’#happy’, ’#love’, ’#surprised’, etc. To create a dataset of tweets rich in a particular emotion, they have used the following methodology. For each emotion X, they selected 50 to 100 terms that were associated with that emotion at different intensity levels. For example, the angry dataset used these terms as follows mad, frustrated, annoyed, peeved, irritated, miffed, fury, antagonism, and so on. This dataset consists of 4 emotion class anger, fear, joy and sadness, they have represented anger and disgust as anger and happiness and sadness as joy.

The dataset of the task was divided into 3 languages as follows English, Arabic and Spanish. In each language there are 5 sub-task datasets. We only work with EI-oc subtask dataset. In which for each tweet there is an emotion alongside the corresponding intensity of that tweet [3]. An initial distribution of the dataset’s task EI-oc can be found in figure 2.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

PROPOSED SYSTEM :

Raw tweets scraped from twitter usually results in a noisy dataset with a lot of useless value. This is due to the nature of the user's usage of Twitter in their own way. Tweets have certain exceptional characteristics such as website URL, short form words, retweets, emoticons, person mentions, etc. which have to be suitably extracted [17]. Therefore, raw twitter data has to be pre-processed to create a dataset which will be easy for different classifiers to generate good results. We have utilized a great variety of pre-processing steps to standardize the dataset and reduce its size. We do the pre-processing on tweets which are as follows.

• Removing tweets that are not in English.

• Converting the tweet to lower case.

• Removing URL from the tweet.

• Removal of mentions, retweet mentions, and unnecessary numbers.

• Separation of hashtags as they can play a vital role in emotion analysis.

• Changing short form words to their full form. E.g. btw stands for by the way.

• Changing the emoticons with their meaning. E.g. (":D") stands for laugh/joy with respect to table I.

• Word tokenizing

• Striping punctuations ['"?!,.():;] from the tokenized words.

• Stop words removal from the tokenized words.

• Stemming and lemmatizing the tokenized words.

• In the end, making parts of speech tag for those tokenized words.

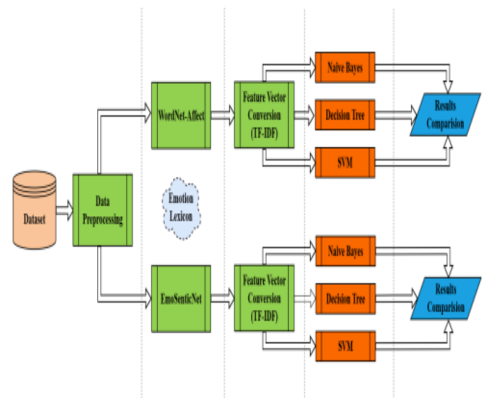


Fig. 1. Proposed Methodology

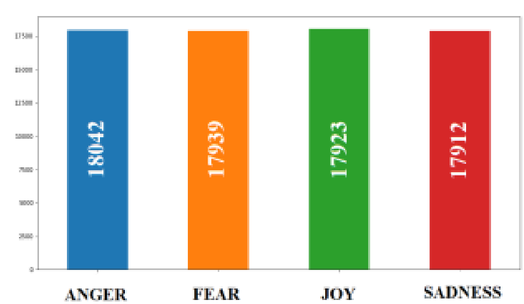


Fig. 2. Visualization of the initial number of tweets in the dataset

TABLE I
EMOJI TO EMOTION KEYWORDS

Emoji	Emotion Keywords
:), :), :-), (:, (:, (-:, :')	Smile
:D, : D, :-D, xD, x-D, XD, X-D	Laugh
<3, :*	Love
;-), ;), ;-D, ;D, (;, (-;	Affection
:-(, : (, :(,):,)-:	Sad
:'(, :'(, :"(	Cry

International Journal For Advanced Research
In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

TABLE II
DATA PREPROCESSING OF TWITTER TWEET

Comment	Tweet
Real Tweet	@SatisfyingTaste @TheAnimalVines I used to make the peanut butter energy balls all the time. My famjam loved them! Btw my cats keep loving them as well :D :D #recipes #yummy 100 ãðŸ https://bit.ly/2orTQLP
Preprocessed Tweet	i used to make the peanut butter energy balls all the time my famjam loved them by the way my cats keep loving them as well recipes yummy joy joy

After this, we have removed the stop words, stemmed using the Porter stemmer which works well for tweet [18] and lemmatized the words and finally POS tagged the words. We can see it’s quiet challenging for us to pre-process the tweet as Twitter itself has its own native way of representation. The most difficult part was to detect the misspelled words and finding the full form of short forms like ASAP is “as soon as possible”. We have used a slang dictionary provided by [19] and using SymSpell [20], we have corrected the misspelled words and compound words.

WORDNET-AFFECT AND EMOSENTICNET

After the data is preprocessed we have used WordNetAffect [14] which is a subset of WordNet with only emotion words. We have mapped those words with our tweet’s words and retrieve the emotion words only (words which represent the emotion of any kind). Then we took those emotions words as features.

EmoSenticNet is another lexical resource that assigns six WordNet Affect emotion labels to SenticNet concept. It can also be thought of as an expansion of WordNet Affect emotion labels to a larger vocabulary [15]. We created a list of emotions words with syntactic relations. After that, we took these as features also.

As the dataset is a labeled one, each tweet has its own emotion with it. But as we are focusing on emotion words in tweets, we filtered out the emotion words from each tweet and stored them into a Pandas DataFrame, which was later used in training and testing the model. Here it should be noticed that we are assuming each tweet has only one emotion words. One example is given below:

Tweet: “Today I’m feeling loved.”

Emotion: “joy”

So we filtered out the emotion word from this tweet.

Emotion word: loved

And store it in the DataFrame with the emotion, Joy. Such as

‘loved’ => ‘Joy’.

The whole process of filtering the tweet with respect WordNet-Affect and EmoSenticNet is shown in Algorithm 1.

FEATURE VECTOR CONVERSION

We represent each tweet into a vector of features for the training of a classifier from labeled data. We must capture features describing each tweet’s emotion. Selection of features plays an important role in the classification process ’ effectiveness.

ess. We employed two well-known techniques to create the feature vector from the 2 features set. These are term frequency (TF) and term frequency and inverse document frequency (TFIDF).

Term Frequency, which measures how frequently a term occurs in a document. Inverse Document Frequency, which measures how important a term is.

$$TF - IDF(T, d, D) = TF(T, d) \times IDF(T, D)$$

Here T is number of terms appeared in a document, d is total number of term in each document, D is the total number of documents.

As after the feature selections we have only the features which bear emotions towards the classification, we have used TF and TF-IDF to increase the importance of those features.

SUPERVISED CLASSIFIER

We have used Naïve Bayesian, Decision Tree and Support Vector Machine in our model.

Bayesian Classifiers build a probabilistic model based on the word features in different classes. For our multi-class emotion classification problem we have taken Multinomial Naïve Bayes Classifier. We have trained our model on MultinomialNB provided by the sk-learn package. In Naive Bayes, texts are classified based on posterior probabilities generated based on the presence of different classes of words in texts. This assumptions makes the computations resources needed for a naïve bayes classifier is far more efficient than non-naïve bayes approaches which is exponential complexity [21]. Naïve Bayes has been widely used for classifying text because it is simple and fast [22], [23], [24].

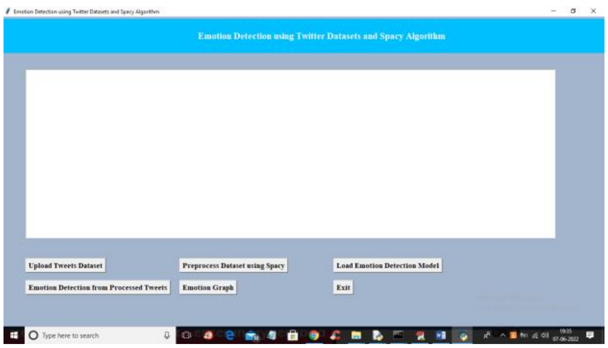
hey accept high dimensional feature spaces and sparse feature vectors. Also, text classification using SVMs is very robust to outliers and does not require any parameter tuning. It finds a maximum margin separating hyper plane between two classes of data. For multi-class classification SVM maximize the margin for one vs all classes of data. For our classification problem we have used Linear SVM of sk-learn package. It has been shown that linear kernels based SVM performs a lot better on non-linear SVM in terms of text classification [21]. Decision trees are slow and sometimes suffer from overfitting. However, its accuracy competes with well-known text classification algorithms such as SVM. For our model we took sk-learn DicisionTreeClassifier. We defined the criterion as entropy as we want the function to measure the quality of a split in text for the information gain. Information gain measures how much organized the input features became after we divide them up using a given feature. Also we have given a static random_state value for our classifier. As for text classification a decision tree takes the features as input values. The decision nodes checks the feature values and leaf nodes which assign one of the classes from our multi class emotion. To choose a class for our input text the model start with the initial node as root node which contains a condition on the input features, it than selects a branch based on that feature which leads to a new condition and it makes a new decision based on it. The

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST ISSN: 2457-0362

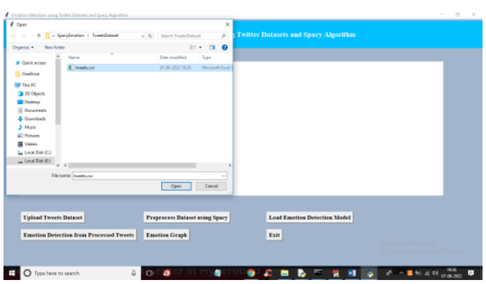
flow continues until it arrives at a leaf node which provides an emotion class for the input value [25].

Results:

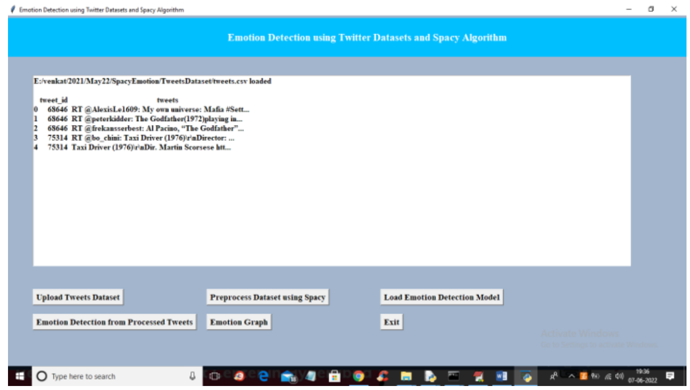
To run project double click on ‘run.bat’ file to get below screen



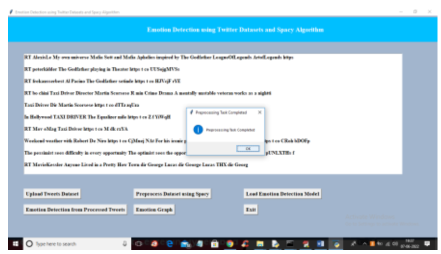
In above screen click on ‘Upload Tweets Dataset’ button to load tweets and get below output



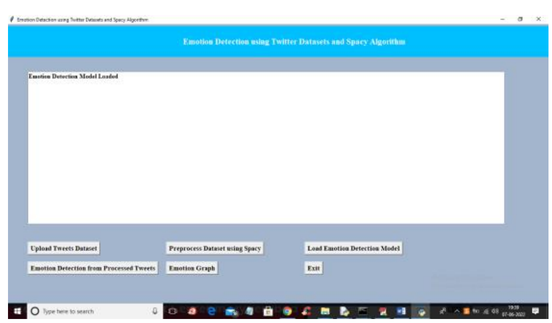
In above screen selecting and uploading tweets dataset and then click on ‘Open’ button to get below output



In above screen we can see dataset loaded and tweets contains total unstructured text with stop words and special symbols and now click on ‘Preprocess Dataset using Spacy’ to clean tweets and get below output

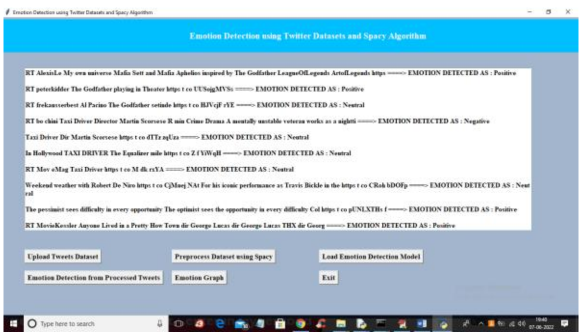


In above screen Preprocessing completed and we can see all tweets contains only text with clean words and now click ‘Ok’ button and then click on ‘Load Emotion Detection Model’ button to load machine learning model for emotion detection and get below output

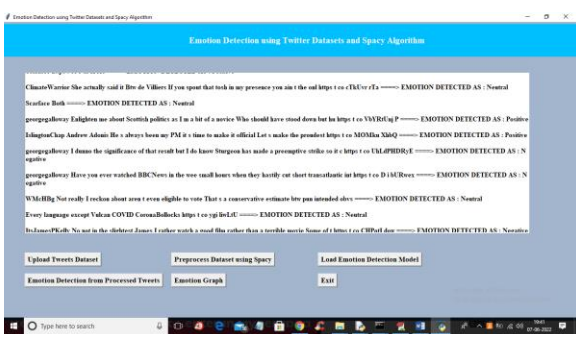


International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

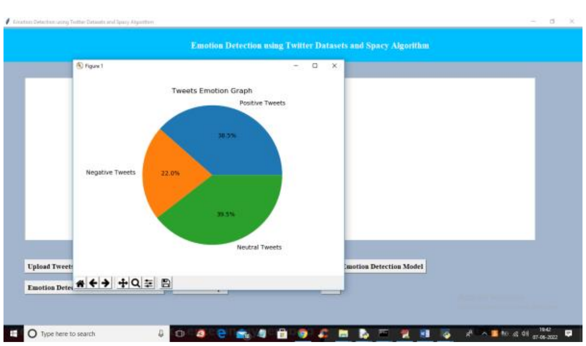
In above screen model is loaded and now click on 'Emotion Detection from Processed Tweets' button to detect emotion and get below output



In above screen before arrow symbol =➔ we can see clean tweet messages and after arrow symbol we can see predicted emotion as 'Positive, Negative or Neutral' and scroll down above screen to view all messages



In above screen we can see all tweets with emotion and now click on 'Emotion Graph' to know tweets percentage in each emotion



In above graph 38.5% peoples are giving positive tweets and 22% gave negative tweets and 39.5% gave neutral tweets so by using this application we can easily extract useful knowledge from peoples reviews weather they are satisfy or not on any topics tweets

CONCLUSION :

Emotion detection is one of the toughest problems to solve. Detecting emotion from text is a challenging work and most of the research works have some kind limitations most importantly, language ambiguity, multiple emotion bearing text, text which does not contain any emotion words etc. Yet we have tried several approaches to detect emotion from twitter. We can say after using EmoSenticNet lexicon, the model performs better than using only WordNet-Affect. It can be also said that our model has performed well but still better results are achievable. As for accuracy, the EmoSenticNet outperforms WordNet-Affect by a great margin. Our limitations are that we have used a small sample as our dataset and there are still language ambiguity problems as we have not been able to address texts which represent multiple emotion at the same time. In the future, we will introduce Deep Learning techniques to identify emotion detection on this dataset.

TABLE VI
ACCURACY OF DIFFERENT CLASSIFIERS

Work	Feature Selection	Classifier	Result
[12]	Logistic Regression	Random Forest	90% on Anger Class
Our Model	EmoSenticNet	SVM	89.28% on Anger Class

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

REFERENCES :

[1] R. Hirat and N. Mittal, “A Survey On Emotion Detection Techniques using Text in Blogposts,” International Bulletin of Mathematical Research, vol. 2, no. 1, pp. 180–187, 2015.

[2] R. Sawyer and G.-M. Chen, “The Impact of Social Media on Intercultural Adaptation,” 2012.

[3] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, “Semeval-2018 Task 1: Affect in Tweets,” in Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 1–17.

[4] R. C. Balabantaray, M. Mohammad, and N. Sharma, “Multi-class Twitter Emotion Classification: A New Approach,” International Journal of Applied Information Systems, vol. 4, no. 1, pp. 48–53, 2012.

[5] M. Anjaria and R. M. R. Guddeti, “Influence factor based opinion mining of Twitter data using supervised learning,” in 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014, pp. 1–8.

[6] S. Yuan, H. Huang, and L. Wu, “Use of Word Clustering to Improve Emotion Recognition from Short Text,” Journal of Computing Science and Engineering, vol. 10, no. 4, pp. 103–110, 2016.

[7] B. Thomas, P. Vinod, and K. A. Dhanya, “Multiclass Emotion Extraction from Sentences,” International Journal of Scientific & Engineering Research, vol. 5, no. 2, 2014.

[8] H. Yang, A. Willis, A. D. Roeck, and B. Nuseibeh, “A Hybrid Model for Automatic Emotion Recognition in Suicide Notes,” Biomedical informatics insights, vol. 5, p. 8948, 2012.

[9] D. T. Ho and T. H. Cao, “A High-order Hidden Markov Model for Emotion Detection from Textual Data,” in Pacific Rim Knowledge Acquisition Workshop, 2012, pp. 94–105.

[10] M. Hasan, E. Rundensteiner, and E. Agu, “Automatic emotion detection in text streams by analyzing Twitter data,” International Journal of Data Science and Analytics, vol. 7, no. 1, pp. 35–51, 2019.

[11] M. Hasan and E. Rundensteiner and E. Agu, “Emotex: Detecting Emotions in Twitter Messages,” 2014.

[12] A. Seyeditabari, S. Levens, C. D. Maestas, S. Shaikh, J. I. Walsh, W. Zadrozny, C. Danis, and O. P. Thompson, “Cross Corpus Emotion Classification Using Survey Data,” This paper was presented at AISB, 2017.

[13] H. Binali, C. Wu, and V. Potdar, “Computational approaches for emotion detection in text,” in 4th IEEE International Conference on Digital Ecosystems and Technologies, 2010, pp. 172–177.

[14] C. Strapparava, A. Valitutti et al., “Wordnet-Affect: an Affective Extension of WordNet,” in Lrec, vol. 4, 2004, p. 40.

[15] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, and S. Bandyopadhyay, "Enhanced SenticNet with Affective Labels for ConceptBased Opinion Mining," IEEE Intelligent Systems, vol. 28, no. 2, pp. 31–38, 2013.

[16] E. Cambria, D. Olsher, and D. Rajagopal, "SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis," in Twenty-eighth AAAI conference on artificial intelligence, 2014.

[17] D. Effrosynidis, S. Symeonidis, and A. Arampatzis, "A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis," in International Conference on Theory and Practice of Digital Libraries, 2017, pp. 394–406.

[18] A. G. Jivani et al., "A Comparative Study of Stemming Algorithms," Int. J. Comp. Tech. Appl, vol. 2, no. 6, pp. 1930–1938, 2011.

[19] "Slang_Dict," https://floatcode.files.wordpress.com/2015/11/slang_dict. doc, accessed: 2019-05-17.

[20] "Github - wolfgarbe/SymSpell," https://github.com/wolfgarbe/SymSpell, accessed: 2019-05-17.

[21] Y. Yang, X. Liu et al., "A re-examination of text categorization methods," in Sigir, vol. 99, no. 8, 1999, p. 99.

[22] H. Zhang, "The optimality of naive bayes," AA, vol. 1, no. 2, p. 3, 2004.

[23] G. Forman and I. Cohen, "Learning from little: Comparison of classifiers given little training," in European Conference on Principles of Data Mining and Knowledge Discovery. Springer, 2004, pp. 161–172.

[24] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in Advances in neural information processing systems, 2002, pp. 841–848.

[25] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python, 1st ed. O'Reilly Media, Inc., 2009.